

Практическая работа № 1

Дискретный вариационный ряд и его характеристики

1. Построение дискретного вариационного ряда

Почти все встречающиеся в жизни величины (урожайность сельскохозяйственных растений, продуктивность скота, производительность труда и заработная плата рабочих, объем производства продукции и т.д. и т.п.) принимают неодинаковые значения у различных членов совокупности. Поэтому возникает необходимость в изучении их изменчивости. Это изучение начинается с проведения соответствующих наблюдений, обследований.

В результате наблюдений получают сведения о численной величине изучаемого признака у каждого члена данной совокупности.

Пример 1.1. Предположим, что мы, интересуясь обеспеченностью хозяйств области зерноуборочными комбайнами Дон-1500, получили следующие данные (в порядке их поступления):

1	3	4	4	6	5	5	2	2	3
3	5	5	2	1	6	3	1	3	3
1	5	2	3	2	2	4	5	1	6
6	4	3	2	4	2	2	3	2	4
4	2	3	4	2	2	2	2	2	3
3	2	3	4	3	2	3	3	2	2

Как видим, интересующий нас признак принимает только целые значения. Он меняется от одного члена совокупности к другому, варьирует. Итак, *варьирование* есть изменчивость признака у отдельных членов совокупности.

Для построения дискретного вариационного ряда составим таблицу.

Таблица 1.1

Распределение 60 хозяйств области по обеспеченности зерноуборочными комбайнами Дон-1500.

Кол-во комбайнов, шт. (x_i)	1	2	3	4	5	6	Итого
Число хозяйств (m_i)	5	20	16	9	6	4	60

Определение 1.1. Вариационным рядом называется последовательность вариантов, записанных в возрастающем порядке и соответствующих частот.

Определение 1.2. Число, показывающее, сколько раз повторяется в данной совокупности каждое значение признака, называется частотой.

При большом количестве вариантов значений дискретного признака может возникнуть ситуация, когда многие из них будут иметь небольшие частоты. В результате ряд будет недостаточно полно отражать общую закономерность распределения признака в совокупности, поскольку на фактическое соотношение частот здесь в значительной мере влияют случайные факторы. В таких случаях производится укрупнение групп. Формально это может означать переход к интервальному ряду распределения, порядок построения которого будет рассмотрен в практической работе №2.

2. Графическое изображение вариационных рядов

Геометрическая иллюстрация статистических данных, геометрическая интерпретация отдельных вопросов статистики придает им наглядность, а в ряде случаев позволяет подвергнуть их анализу в наиболее простой и доступной форме. Это в полной мере относится и к графическому изображению вариационных рядов.

Применяется несколько способов графического изображения рядов распределений в зависимости от вида их и от поставленной задачи: полигон, гистограмма, кумулятивная кривая (кумулята), огива.

Определение 1.3. Полигоном частот называют ломаную, отрезки которой соединяют точки $(x_1; m_1), (x_2; m_2), \dots, (x_k; m_k)$, где x_i - варианты выборки и m_i - соответствующие им частоты.

Определение 1.4. Отношения соответствующих частот к объему совокупности называются *относительными частотами (частостями)*.

Относительные частоты (частости) обозначаются w_i , и находятся по формуле:

$$w_i = \frac{m_i}{n}$$

Определение 1.5. Полигоном относительных частот называют ломаную, отрезки которой соединяют точки $(x_1; w_1), (x_2; w_2), \dots, (x_k; w_k)$, где x_i - варианты выборки и w_i - соответствующие им относительные частоты.

Гистограмма частот (гистограмма относительных частот) строится в случае непрерывной вариации (для интервального ряда).

Определение 1.6. Накопленной частотой (частостью) называют сумму всех частот (частостей) вариационного ряда, предшествующих данной варианту с частотой (частостью) этой варианты.

Если накопленную частоту варианты x_i , обозначить S_i , то ее можно будет найти по формуле:

$$S_i = \sum_{k=1}^i m_k$$

Аналогично можно будет найти накопленную частоту (относительную частоту), разделив накопленную частоту на объем совокупности n .

Определение 1.7. *Кумулятивной кривой* называется ломаная, отрезки которой соединяют точки $(x_1; S_1), (x_2; S_2), \dots, (x_k; S_k)$, где x_i - варианты выборки и S_i - соответствующие им накопленные частоты.

Если на горизонтальной оси откладывать накопленные частоты, а на вертикальной – значения признака, то полученная при этом кривая носит название *огивы*.

Таким образом, по сравнению с кумулятивной кривой при построении огивы оси абсцисс и ординат меняются ролями. Если построенные по одним данным и с одинаковым масштабом графики кумуляты и огивы наложить один на другой, то они будут симметричными относительно прямой, проведенной под углом 45 градусов из начала координат. Таким образом, принципиальных различий между кумулятой и огивой нет, оба графика выполняют одинаковые функции.

Замечание. Кумуляту и огиву, как, впрочем, и полигон, можно построить не только для дискретных, но и для интервальных рядов. Да и гистограмму можно построить не только для интервальных, но и для дискретных рядов.

➤ Для построения графиков составим вспомогательную таблицу.

Таблица 1.2.

Кол-во ком-байнов, шт. (x)	1	2	3	4	5	6	Итого
Число хозяйств (m_i)	5	20	16	9	6	4	60
Накопленная частота (S_i)	5	25	41	50	56	60	
Относительная накопленная частота $\left(\frac{S_i}{n}\right)$	0,08	0,42	0,68	0,83	0,93	1	

Полигон частот



Полигон частот(замкнутый)



Замечание. Чтобы полигон получился замкнутым, необходимо дополнить рабочую таблицу двумя крайними значениями, которые являются ближайшими к имеющимся в ней значениями из числа возможных. Такими вариантами являются $x_0 = 0$ и $x_7 = 7$. Их сейчас нет в рабочей таблице лишь потому, что среди хозяйств области нет таких, в которых комбайнов Дон-1500 не было совсем или же было 7 штук. Этим вариантам соответствуют равные нулю частоты. Аналогичным образом можно поступить с кумулятой и огивой.

Кумулята



Огива



3. Эмпирическая функция распределения

Определение 1.8. Эмпирической функцией распределения называется такая функция, которая для каждого x выражает долю (частость) тех наблюдений, в которых рассматриваемый признак принял какое-нибудь значение, меньшее x , т.е.

$$F_n(x) = \frac{n(x)}{n}$$

где $n(x)$ - число вариантов, в которых рассматриваемый признак принял какое-нибудь значение, меньшее x ; n - общее число наблюдений (объем выборки).

Таким образом, $n(x)$ соответствует накопленной частоте предыдущей варианты.

- Найдем эмпирическую функцию распределения по количеству комбайнов Дон-1500 в хозяйствах области.

Искомая эмпирическая функция распределения равна нулю для всех значений x от $-\infty$ до 1 включительно. Покажем сначала, что она равна нулю при $x = 1$. По определению

$$F_n(1) = \frac{n(1)}{60}$$

где $n(1)$ означает число хозяйств области, в которых меньше 1 комбайна. В рассматриваемом примере таких нет, т.е. $n(1) = 0$. Отсюда вытекает, что $F_n(1) = 0$. Тогда тем более $F_n(x) = 0$ для значений $x < 0$, так как число комбайнов в хозяйствах не может быть отрицательным.

Пусть теперь x удовлетворяет неравенствам $1 < x \leq 2$, т.е. от первой до второй варианты в таблице 1.2. Так как дробное значение мы взять не можем, в силу того, что число комбайнов является целым числом, проверим случай, когда $x = 2$. Подсчитаем, которое равно числу хозяйств области, в которых меньше двух комбайнов. Таких хозяйств, согласно таблице 1.2, пять. Найдем теперь $F_n(2) = \frac{5}{60} = 0,08$.

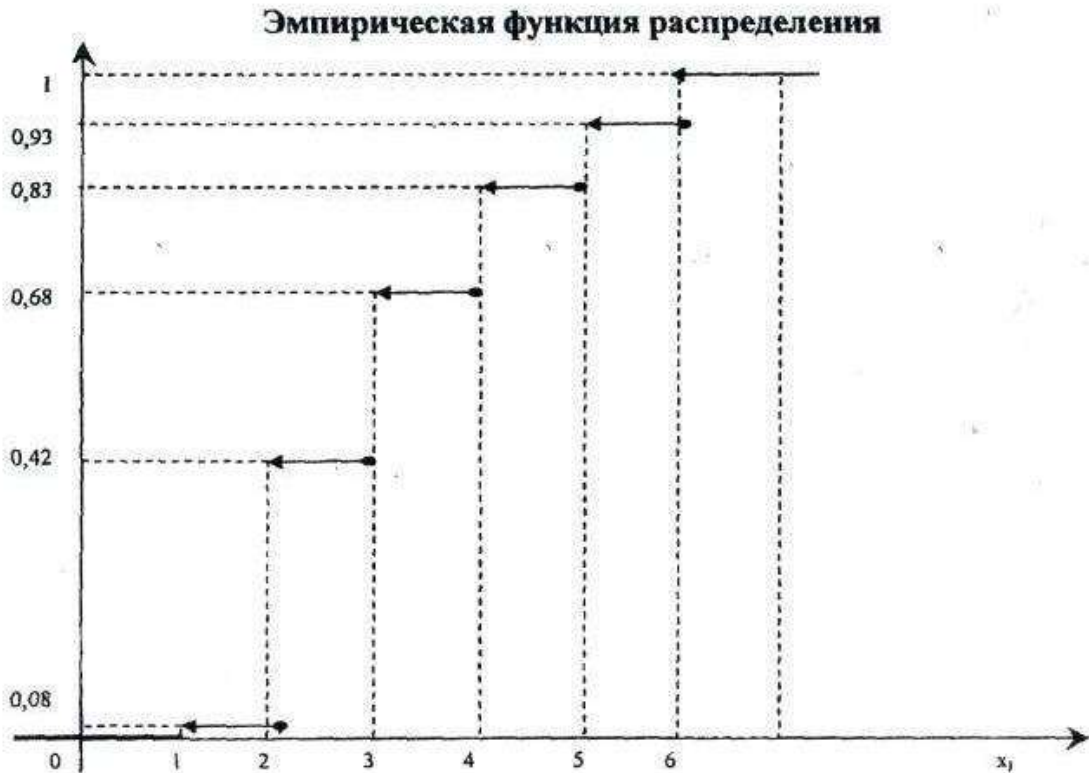
Предположим далее, что x удовлетворяет неравенствам $2 < x \leq 3$, т.е. от второй до третьей варианты таблицы 1.2. Возьмем $x = 3$. Подсчитаем $n(3)$ которое равно числу хозяйств области, в которых меньше трех комбайнов. Для этого нам нужно сложить число хозяйств, в которых один комбайн и число хозяйств, в которых два комбайна, т.е. найти накопленные частоты. Таких хозяйств, согласно таблице 1.2, двадцать пять. Найдем теперь $F_n(3) = \frac{25}{60} = 0,42$.

Аналогично рассуждая, можно найти значение искомой эмпирической функции распределения для любого значения x . При этом можно воспользоваться результатом нахождения относительных накопленных частот табл. 1.2.

В результате получим следующее выражение искомой эмпирической функции распределения:

$$F_n(x) = \begin{cases} 0, & \text{при } x \leq 1, \\ 0,08 & \text{при } 1 < x \leq 2, \\ 0,42 & \text{при } 2 < x \leq 3, \\ 0,68 & \text{при } 3 < x \leq 4, \\ 0,83 & \text{при } 4 < x \leq 5, \\ 0,93 & \text{при } 5 < x \leq 6 \\ 1 & \text{при } x > 6. \end{cases}$$

Изобразим эту функцию графически:



4. Числовые характеристики выборки

Находим:

1. выборочную среднюю \bar{x} по формуле: $\bar{x} = \frac{\sum_{i=1}^k x_i m_i}{n}$,

где x_i - варианты, m_i - соответствующая частота, k - количество различных вариантов, n - объем выборки.

2. выборочную дисперсию $D(X)$ по формуле: $D(X) = \frac{\sum (x_i - \bar{x})^2 m_i}{n}$.

3. «исправленную» дисперсию: $s^2 = \frac{n}{n-1} \cdot D(X)$.

4. «исправленное» среднее квадратическое отклонение: s

Таблица 1.3.

Вспомогательная таблица для расчета числовых характеристик ряда распределения

Кол-во ком-байнов, шт, x_i	Число хо-зяйств (m_i)	$x_i m_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 m_i$
1	5	5	-2,05	4,2025	21,0125
2	20	40	-1,05	1,1025	22,05
3	16	48	-0,05	0,0025	0,04

Продолжение таблицы 1.3

4	9	36	0,95	0,9025	8,1225
5	6	30	1,95	3,8025	22,815
6	4	24	2,95	8,7025	34,81
Σ	60	183	-	-	108,85
$\frac{\Sigma}{60}$	-	$\bar{x} = 3,05$	-	-	$D(X) = 1,81$

«Исправим» дисперсию: $s^2 = \frac{n}{n-1} \cdot D(X) = \frac{60}{59} \cdot 1,81 = 1,84$.

Найдем «исправленное» среднее квадратическое отклонение:
 $s = \sqrt{1,84} = 1,35$.

5. Мода

Определение 1.9. Модой M_0 называется наиболее часто встречающаяся варианта.

Нахождение моды дискретного распределения не требует каких-либо вычислений - ею является варианта, которой соответствует наибольшая частота. В рассматриваемом примере $M_0 = 2$.

6. Медиана

Определение 1.10. Медианой M_e называется варианта, приходящаяся на середину вариационного ряда.

Иными словами, медианой является варианта, делящая совокупность на две равные по объему части. До медианы и после нее имеется одинаковое число членов совокупности. При нахождении медианы дискретного вариационного ряда следует различать два случая:

- 1) объем совокупности нечетный;
- 2) объем совокупности четный.

Пусть объем совокупности нечетный и равен $2n+1$. Расположим все варианты $x_1, x_2, \dots, x_{n-1}, x_n, x_{n+1}, \dots, x_{2n+1}$ в возрастающем порядке. В этом ряду каждая варианта повторена столько раз, сколько она встречается в совокупности. Поэтому среди них могут быть и одинаковые. Медианой этого распределения является вариант с номером n , так как он находится в середине ряда:

$$M_e(X) = x_{n+1}$$

Если объем совокупности четный - равен $2n$, то в ряду $x_1, x_2, \dots, x_n, x_{n+1}, \dots, x_{2n}$ нет варианты, которая делила бы совокупность на две равные по объему части. Поэтому за медиану условно принимают полусумму находящихся в середине ряда вариант. Ими являются варианты с номерами n и

$n+1$:

$$M_e(X) = \frac{x_n + x_{n+1}}{2}$$

➤ В рассматриваемом примере четное число вариант $2n = 60$.

Для нахождения медианы используем таблицу 1.2, данные по накопленным частотам. В таблице находим последнюю накопленную частоту, которая не превосходит половины объема совокупности, и первую, которая больше ее. Они равны соответственно 25 и 41 и показывают, что первые 25 вариант принимают значения, меньшие 3 шт, а следующие 16 вариант с номерами от 26-го до 41-й включительно, в том числе и 30-й и 31-й, принимают значение 3 шт. Следовательно, медиана данного распределения равна:

$$M_e(X) = \frac{3+3}{2} = 3.$$

Медиана может быть определена и графически по кумуляте или огиве. Для определения медианы по кумуляте последнюю ординату, пропорциональную сумме всех частот или частостей, делят пополам. Из полученной точки восстанавливают перпендикуляр до пересечения с кумулятой. Абсцисса точки пересечения и дает значение медианы.

➤ Определим по графику медиану.

Последняя ордината, как видно из графика, равна 60. Деление этой ординаты пополам дает точку А(30). Перпендикуляр из точки А до пересечения с кумулятой дает точку В. Абсцисса точки В, округленная до целых и будет медианой.

Кумулята



Как видно из графика кумуляты, $M_e = 2,8 \approx 3$, что соответствует медиане полученной аналитически.

Проверка гипотезы о нормальном распределении генеральной совокупности по критерию Пирсона

1. Дискретный вариационный ряд

Для того чтобы при заданном уровне значимости α проверить гипотезу о нормальном распределении генеральной совокупности нужно, учитывая ранее вычисленные значения \bar{x} и σ , вычислить теоретические частоты, сравнить эмпирические и теоретические частоты с помощью критерия Пирсона.

1.1. Вычисление теоретических частот

Используем данные из практической работы №1: $\bar{x} = 3,05$ $\sigma = 1,35$
 Вычислим теоретические частоты по формуле:

$$m_i' = \frac{n \cdot h}{\sigma} \cdot \varphi(u_i),$$

где n - объем выборки, h - шаг, $u_i = \frac{x_i - \bar{x}}{\sigma}$, $\varphi(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}$.

Замечания.

1. Функция $\varphi(u)$ является локальной функцией Лапласа, значения которой можно найти по таблицам приложения 1.
2. Функция $\varphi(u)$ является четной, т.е. $\varphi(-u) = \varphi(u)$.

Вычислим величину $\frac{n \cdot h}{\sigma} = \frac{60 \cdot 1}{1,35} = 44,60$ и заполним таблицу.

Таблица 3.1.

Номер варианты, i	Варианта, x_i	«Нормированная варианта», $u_i = \frac{x_i - \bar{x}}{\sigma} = \frac{x_i - 3,05}{1,35}$	Значение функции вероятностей «нормированной варианты», $\varphi(u_i)$	Теоретическая частота, $44,6 \cdot \varphi(u_i)$
1	1	-1,52	0,1257	5,61
2	2	-0,78	0,2943	13,13
3	3	-0,04	0,3986	17,78
4	4	0,71	0,3101	13,83
5	5	1,45	0,1394	6,22
6	6	2,19	0,0363	1,62
Σ				58,17

Замечание. Сумма теоретических частот оказалась отличной от той суммы эмпирических частот. На это расхождение, связанное с погрешностями вычислений.

Вполне допустимо, т.к. составляет менее 5%: $\Delta = \frac{60 - 58,17}{60} \cdot 100\% = 3,05\%$.

1.2. Нахождение наблюдаемого значения критерия $\chi^2_{набл}$

Наблюдаемое значение критерия $\chi^2_{набл}$ вычисляется по формуле:

$$\chi^2_{набл} = \sum \frac{(m_i - m_i')^2}{m_i'}$$

где m_i' - теоретические частоты,

m_i - соответствующие эмпирические частоты.

Наблюдаемое значение критерия найдем с помощью таблицы.

Таблица 3.2.

Номер варианты, i	Эмпирические частоты, m_i	Теоретические частоты, m_i'	$m_i - m_i'$	$(m_i - m_i')^2$	$\frac{(m_i - m_i')^2}{m_i'}$
1	5	5,61	-0,61	0,3721	0,0664
2	20	13,13	6,87	47,1969	3,5959
3	16	17,78	-1,78	3,1684	0,1782
4	9	13,83	-4,83	23,3289	1,6869
5	6	6,22	-0,22	0,0484	0,0078
6	4	1,62	2,38	5,6644	3,4965
Σ	60	58,17			9,0317

Итак, мы нашли наблюдаемое значение критерия $\chi_{\text{набл}}^2 = 9,0317$.

1.3. Нахождение критических точек распределения χ^2

Найдем число степеней свободы k по формуле:

$$k = s - r - 1,$$

где s - число групп выборки, r - число параметров, оцениваемых по выборке.

Нормальное распределение оценивается двумя параметрами: математическим ожиданием a и средним квадратическим отклонением σ . Так как оба эти параметра оценивались по выборке (в качестве оценки a принимают выборочную среднюю, в качестве оценки σ - выборочное среднее квадратическое отклонение), то $r=2$. Поэтому

$$k = 6 - 2 - 1 = 3.$$

По таблице критических точек распределения χ^2 (приложение 3), по заданному уровню значимости $\alpha = 0,05$ и числу степеней свободы $k = 3$ находим критическую точку $\chi_{\text{табл}}^2(\alpha, k)$:

$$\chi_{\text{табл}}^2(0,05;3) = 7,8.$$

Так как $\chi_{\text{набл}}^2 > \chi_{\text{табл}}^2$ - гипотезу о нормальном распределении генеральной совокупности отвергаем. Эмпирические и теоретические частоты различаются значимо.